**Erklärbarkeit in der Bildklassifikation – Entwicklung eines interaktiven Visualisierungstools**

**Zielsetzung**

Deep-Learning-Modelle erzielen in der Bildklassifikation hervorragende Ergebnisse, gelten jedoch oft als „Black Boxes“. Anwendenden fällt es schwer nachzuvollziehen, warum ein Modell eine bestimmte Entscheidung trifft. Ziel dieses Projekts ist die Entwicklung eines interaktiven Visualisierungstools, das Nutzenden die Möglichkeit gibt, Modellentscheidungen transparent und verständlich nachzuvollziehen.

Im Fokus stehen Methoden der erklärbaren Künstlichen Intelligenz (XAI) [1, 2], die visuelle Erklärungen generieren, wie etwa Grad-CAM [3], Layer-wise Relevance Propagation (LRP) [4] oder Integrated Gradients [5]. Das Tool soll verschiedene XAI-Methoden vergleichbar machen und die Ergebnisse in einer benutzerfreundlichen Oberfläche darstellen.

**geplante Vorgehensweise**

1. **Literatur- und Methodenrecherche** zu XAI-Ansätzen für Convolutional Neural Networks (CNNs).
2. **Implementierung** ausgewählter Verfahren (z. B. Grad-CAM, LRP) für gängige Bildklassifikationsmodelle (ResNet [6] oder EfficientNet[7]).
3. **Entwicklung einer interaktiven Visualisierungsoberfläche**
4. **Evaluation** der Erklärungen anhand von Fallstudien
5. **Nutzerstudie (optional):** Untersuchung, wie verständlich und hilfreich die generierten Erklärungen für verschiedene Zielgruppen sind.

**Erwartete Ergebnisse**

- Vergleich verschiedener XAI-Methoden im Hinblick auf Verständlichkeit und Aussagekraft.
- Funktionsfähiges Prototyp-Tool zur Visualisierung von Modellentscheidungen.
- Diskussion von Chancen und Grenzen der visuellen Erklärbarkeit.

**Benötigte Voraussetzungen für Studierende**

- Kenntnisse in **Python** und gängigen ML-/DL-Frameworks (z. B. PyTorch oder TensorFlow).
- Grundlagenwissen in **Deep Learning**, insbesondere CNNs.
- Basiskenntnisse in **Webentwicklung oder GUI-Design** sind von Vorteil, aber nicht zwingend erforderlich.

**Referenzen**


1. Molnar, C.: Interpretable machine learning: a guide for making black box models explainable. Christoph Molnar, Munich, Germany (2022).
2. Vilone, G., Longo, L.: Classification of Explainable Artificial Intelligence Methods through Their Output Formats. Machine Learning and Knowledge Extraction. 3, 615–661 (2021). https://doi.org/10.3390/make3030032.
3. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: 2017 IEEE

International Conference on Computer Vision (ICCV). pp. 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74.
4. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLOS ONE. 10, e0130140 (2015). https://doi.org/10.1371/journal.pone.0130140.
5. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic Attribution for Deep Networks, http://arxiv.org/abs/1703.01365, (2017). https://doi.org/10.48550/arXiv.1703.01365.
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778. IEEE, Las Vegas, NV, USA (2016). https://doi.org/10.1109/CVPR.2016.90.
7. Tan, M., Le, Q.V.: EfficientNetV2: Smaller Models and Faster Training. (2021). https://doi.org/10.48550/ARXIV.2104.00298.